A Context-Aware Recall Survey for Data Collection Using Ubiquitous Sensors in the Home

ABSTRACT

Identifying what people do in the home can both inform ubiquitous computing application design decisions and provide training data to the machine learning algorithms used in their implementation. Ideally, data collection in the home should be non-invasive, automatic and inexpensive, while data labeling should be scalable depending on user abilities. This paper describes an unsupervised technique in which contextual information gathered by ubiquitous sensors is used to help users label a multitude of anonymous activity episodes. This *context-aware* recall survey is a game-like computer program in which users attempt to correctly guess which activity is happening after seeing a series of symbolic images that represent sensor values generated during the activity. We report a user study of the system, focusing on how well subjects were able to recognize their own activities, the activities of others, and counterfeits that did not correspond to any activity. We present a series of "lessons learned" and describe future work incorporating these findings.

Author Keywords

Ubiquitous computing, context-aware computing, home, activity recognition, recall survey, experience sampling.

ACM Classification Keywords

I.2.6 **[Learning]:** Knowledge Acquisition. H.5.m **[HCI]:** Miscellaneous. H.3.m **[Information Retrieval]:** Miscellaneous.

INTRODUCTION

Many ubiquitous computing applications depend on knowledge of how people behave in their environments. Labeled examples of activity can improve design decisions and help machine learning techniques to recognize and predict activity. However, collecting this information can be a delicate process (especially in the home). The work in this paper is motivated by a need to acquire labeled examples of activity in the home without burdening users or disrupting their daily routines.

Several standard classes of methods exist for collecting data about daily activities, including one-on-one or group interviews, direct observation, self report recall surveys, time diaries, and the experience sampling method (ESM) [2, 5]. While direct observation is often reliable, it is prohibitively time-consuming. In interviews and recall surveys, users often have trouble remembering activities and may censor what they do report. Cognitively enhanced recall surveys mitigate forgetfulness by using cues such as photo snapshots. Time diaries also reduce recall and selective reporting bias, but require a commitment from the user to carry around (and use) the diary. Experience sampling uses a prompting mechanism (e.g., a beep) to periodically "ask" the user for a self-report. These prompts may interrupt activities and must be carefully delivered in order to avoid annoying the user [5]. All of these methods require the participation of the person who performed the activity and others may require outside help as well (e.g., interviewers).

In this paper we describe a *context-aware* recall survey (CARS) that is designed to allow users to label activities via an augmented recall survey that displays contextual information collected by ubiquitous sensors. Drawing on recent research in practical home monitoring systems and game-based image-labeling techniques [1, 8], we designed a game-like multiple choice test that converts low-level sensor readings into colorful symbols and descriptive text. Users answer the questions with the goal of correctly labeling the activity being depicted. This approach allows anyone to label the data at any time, without requiring additional hardware (beyond sensors) or causing additional interruption to daily routine.

We report a study in which users (N=10) performed a subset of tasks in an instrumented environment and completed a CARS approximately one week later. The results draw upon a statistical analysis of test performance and a post-test questionnaire. We present a comparative analysis of how well users performed on their own activities (which they may remember) vs. the activities of others (which they have never seen) vs. counterfeit examples of activity. We summarize our results with a series of lessons learned. These findings have value that extends beyond the specifics of our own system, both in terms of the design of data collection systems and in terms of research on the abilities and needs of human labelers. Our method is not necessarily meant to replace other methods, but to provide a stronger solution when used with existing approaches.

BACKGROUND

We discuss three background topics; each motivates a different aspect of the system.

The "Box of Sensors" Scenario

Existing sensor infrastructure is increasingly prevalent in the home. The cost of sensors and computation has dropped to the point that consumer-ready sensor installations are becoming possible. A study conducted by Intel Research Seattle found that participants were able to install faux sensors according to a simple instruction sheet, without expert supervision [4]. Meanwhile, other researchers are designing and deploying real sensor installations in the home [3, 8].

We envision the CARS as a tool usable by non-experts to "train" their in-home monitoring systems. Simple, do-it-yourself sensors are likely to constitute the first examples of actual in-home monitoring systems. These sensors are chosen specifically for real homes; they are low-cost, affordable, and not perceived as invasive. We have taken the following as a design challenge: to use output from an existing sensor infrastructure to help users label activities without any additional impact to daily routines.

Context-Aware Experience Sampling Method

Existing data collection schemes increasingly exploit an underlying sensor infrastructure. The electronic experience sampling method (ESM) collects data by using a portable computing device (e.g. a PDA or cell phone) to prompt users and to collect their responses. Prompting can interrupt activities and quickly become an annoyance [2]. In response, *context-aware* ESM approaches use additional sensor information to choose the best time to prompt for data [6].

Our goal is for CARS techniques to mesh seamlessly with context-aware ESM techniques. Two immediate benefits emerge: 1) CARS could be used to "fill in the blanks" from missed or ignored ESM prompts, and 2) labels collected via ESM could be verified or supplemented.

Use of Online Games for Labeling

The human desire to be entertained can be tapped for labeling data. Researchers at CMU recently showed the potential of using "human cycles" to label images on the Internet [1]. In the "ESP Game," two online players examine the same image and independently type descriptive words. The result is a double-blind test that rewards players who come up with the same answers.

Our method currently engages a single user off-line with a scrolling series of colorful, descriptive images and then asks for the most likely activity classification. This setup lends itself well to the transition to an online-game. The advantage is at the least to provide a second opinion and at most to provide a label when it is infeasible (or inconvenient) to ask the performer of the activity for it.

CONTEXT-AWARE RECALL SURVEY

The key idea of a CARS is to use contextual information collected by ubiquitous sensors to provide an improved recall survey that can be performed by anyone at any time, regardless of who performed the activity or how the sensors were configured. After sensor readings are collected, the following steps occur: 1) sensor readings are automatically segmented by activity into episodes, 2) episodes are converted into a series of generic, highly descriptive images, and 3) episodes are labeled by users in a game-like computer-based recall survey.

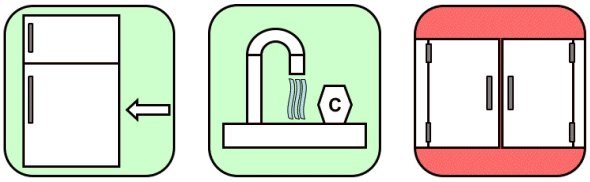

Figure 1. (a) Fridge open, (b) water on, (c) cabinet close.

Unsupervised segmenting of time-series sensor data into episodes is known as the *episode recovery* problem. In the study described in this paper, we performed episode recovery by hand. Unsupervised methods are more desirable because they require no human intervention, but they are also more error prone. Our study attempts to address the implications of malformed episodes by asking users to label *counterfeit episodes*. Counterfeit episodes were hand-made by intentionally choosing incorrect activity durations and by introducing noisy sensor readings (i.e., dropped readings and spurious readings).

An unlabeled episode is translated into a symbolic image and short length of text via a tool called the Narrator [7]. The Narrator is a finite state machine that parses low-level sensor information and generates a concise, readable summary. For this study, a simplified version of the Narrator mapped each change in sensor values to an image and caption. See Figure 1 for an example. The source of the sensor reading (e.g., contact switch, motion detector, break-beam sensor) is inconsequential, so long as it can be mapped to a symbol. Each new sensor configuration must be hand-mapped to a standard set of generic images that may be shared by many other instrumented homes.

The resulting series of images (and captions) describes an activity episode in simple, symbolic terms. These episodes are then presented to users in the form of a game-like recall survey. We describe the particular implementation of our CARS in the next section.

STUDY

Subjects. Subjects in this study were 10 adult volunteers who were recruited from the university and from the community. Subjects ranged in age from 25 to 32 years, and the sample was 50% female and 50% male. Subject background varied, ranging from librarians to engineers.

Instrumented environment. This study occurred in the lead author's home. A kitchen and bathroom were instrumented with two types of anonymous, binary sensors: magnetic contact switches and pressure mats. In the kitchen, contact swit-

ches were installed on the refrigerator, freezer, and microwave doors, as well as on the trash can, liquid soap dispenser, stove top burner, hot and cold water faucet knobs, two cupboards, and two drawers. A pressure mat was placed under the chair at the kitchen table. The refrigerator, freezer, cupboards and drawers used in the experiment were clearly labeled with their contents (e.g., ice, plates, cups). In the bathroom, magnetic contact switches were installed on the liquid soap dispenser and hot and cold water faucet knobs, and a pressure mat was placed on the floor in front of the sink. All sensors interfaced with a desktop computer via an extended parallel port. Sensors were polled every second and values were stored in a mySQL database.

Activity recording. Subjects participated one at a time. First, subjects were asked to perform several activities in the instrumented kitchen and bathroom. They were informed that they would be asked to recall their activities in a "quiz" later that week. The locations of objects needed to perform the activities were clearly labeled and each subject was given an initial tour of the instrumented rooms. Subjects were then instructed to choose and perform a subset of several kitchen tasks (which were also posted on the refrigerator as a reminder). The kitchen tasks were: prepare a cold drink, prepare either a sandwich, a fried egg, or a microwave pizza, eat the meal, wash dishes and put them away, and throw away any trash. During the bathroom portion, subjects were given a toothbrush and were instructed to brush their teeth and then perform two of three tasks: washing their face, washing their hands, and combing their hair. An observer time-stamped the start and end points of each activity using a laptop computer. Subjects were instructed not to speak with other subjects about which tasks they had performed.

Context Aware Recall Survey. We presented our computer-based recall survey as a "game" in which the goal was to correctly guess which activities were happening given only the sensor readings collected from the kitchen and bathroom environments. The contextual information gathered by the sensors was hand-segmented and converted into episodes of images and text via the Narrator program [7]. Several counterfeit episodes (which did not correspond to any activity) were generated by hand. The rest of the episodes were "real" in that they were generated by subjects performing activities in the instrumented environment.

Each episode consisted of a series of scrolling images that had red or green backgrounds, depending on whether that object was turned on or off. The word "kitchen" or "bathroom" was presented with each episode to indicate the location of the episode. The total duration of the episode was also displayed (no other timing information was included). Subjects were able to pause the scrolling pictures, but were not able to replay an episode. After viewing an episode, subjects were asked to select from a multiple choice list of every possible kitchen or bathroom activity (depending on which room the activity occurred in) plus a "None of the Above" answer. Subjects were also asked to rate how confident they were about their choice on a scale of one to five.

Subjects were administered the CARS on a laptop computer a mean of 5 days following the activity recording. Each subject was presented with two sets of 12 activity episodes, which we call the "self" set and the "other" set. The self set contained 8 episodes from the subject's own activities and 4 counterfeit episodes. The other set contained 8 episodes of someone else's activities and 4 counterfeit episodes. Subjects were informed of which sets were self or other. The survey administration was counterbalanced, with half of the subjects presented the self set first, and the other half with the other set first.

Subjective Experience Questionnaire. Immediately after completing the computer-based CARS, subjects completed a brief paper and pencil questionnaire about their subjective experience.

RESULTS

1. Subjects successfully identified 82% of the 24 total episodes (*M* = 19.60, *SD* = 3.47). Assessing confidence in their selection on the Likert scale of 1-5 (1=Not Sure and 5=Very Sure), subjects reported being Mostly Sure (*M* = 3.96, *SD* = 1.03) across all of the episodes. Overall, **subjects were able to successfully label most activities with confidence.**

2. The number of days between activity performance and activity recall ranged from 2 to 7 (M = 5.00, SD = 1.63) and was not significantly correlated with total performance scores, $r(8) = .27, p = .44$. Therefore, we found that **the amount of time between activity and recall did not significantly affect performance in our sample.**

3. Ignoring counterfeit episodes, performance on the self section (*M* = 7.10, *SD* = 1.29) and the other section (*M* = 7.10, *SD* = .99) was identical, with subjects correctly identifying 89% of the 8 possible episodes. There was also no significant difference in subjects' confidence ratings in their identification of their own activities vs. someone elses. **Overall, subjects were equally good at labeling their own or other people's real activities.**

4. Although there were no significant differences in overall performance on self vs. other episode identification, differences did emerge for the most difficult to identify episodes. Due to limited sensor granularity several bathroom tasks (e.g., face washing vs. hand washing) were only recognizable almost solely from memory. The number of errors on these tasks were low in the self section, but more than doubled in the other section. This indicates that **the importance of memory increases for hard-to-recognize activities.**

5. Subjects had better performance on the 16 real episodes (*M* = 14.20, *SD* = 1.47) than on the 8 counterfeit episodes (*M* = 5.40, *SD* = 2.32). This proportional difference (67% for fake, 89% for real) was significant, $t(9) = -2.80, p < .05$. The mean confidence rating was also slightly higher for real (*M* = 3.02, *SD* = .37) vs. counterfeit episodes (*M* = 2.82, *SD* = .43), although this difference was not statistically significant, $t(9) = 1.43, p = .19$. This indicates that although users may be unaware of this deficit, **counterfeit episodes are often mislabeled as real episodes.**

6. The order of test administration (self then other, or vice versa) did not impact overall performance or performance on self vs. other sets, but did impact performance on the identification of counterfeits. Subjects who completed the self section first were significantly better at detecting fake episodes in the other section, $t(8) = 2.36, p < .05$. We hypothesize this is because users become better at spotting counterfeit episodes as they gain more experience. In other words, **practice makes perfect.**

7. Subjects who completed the other section first did not perform significantly better at spotting counterfeits on the self section. This could be because experience gained during the other set did not overcome the inherent gains offered by remembering which activities did or did not occur. This may be because **remembering activities makes it easier to spot counterfeit episodes.** In the follow-up survey one subject wrote "I could remember the steps I took...and I knew it was me."

8. Performance on each episode was significantly related to subjects rating of their own confidence level in their selection for that episode, $r(238) = .16, p < .05$. There was also a significant difference between mean confidence level on correct (*M* = 3.03, *SD* = 1.03) vs. incorrect (*M* = 2.61, *SD* = 1.06) selections, $t(238) = 2.39, p < .01$. This suggests that **user confidence ratings are potentially a useful measure of whether the episode was correctly labeled.**

9. There was no difference found between confidence ratings on self vs. other activity identification, however, in the follow-up survey, subjects reported that in their overall experience of the test it was more difficult to identify someone else's activities than their own $t(9) = -1.95, p < .10$ (significant at the trend level). It appears that **subjectively, subjects feel it is easier to label their own activities.**

10. Subjects reported that their subjective experience of the test was positive. Subjects reported that the symbolic images were "pretty easy" to "very easy" to understand on a Likert scale of 1-5 (*M* = 4.70, *SD* = .48). Open-ended questions that asked what subjects liked and did not like about the CARS also indicated that subjects had a positive experience, with subjects reporting that they liked the color-coding for off and on, and that the images were "cute," "clear," or "easy to understand." All in all, **subjects enjoyed taking the CARS.**

11. Total performance scores ranged from 11 to 23 correct identifications (out of 24 possible), with 90% of the subjects correctly identifying 18 or more of the episodes. Most subjects were fairly adept at labeling, while a few performed significantly worse. Obviously, **not everyone makes a good choice for a labeler.** However, in the follow-up survey several high performing subjects requested the ability to speed up the scrolling. This indicates that when designing the CARS **researchers should plan for the varying abilities of different subjects.**

CONCLUSION

In this paper, we described the *context-aware* recall survey, an approach for labeling activities that uses contextual information collected by sensors. We presented results from a user study that indicate such an approach can be effective. We used our results to offer several lessons learned that could help other researchers design a better CARS.

In future work we plan to conduct a longer-term use study of the system in a whole-house environment. To do so we will make several improvements to the existing system. First, we will introduce automatic episode segmentation so that the only human interaction comes from the labeler. Second, we will convert the existing multiple-choice test into a simple Internet-based game that can be played online. Third, we will extend the range of supported sensors to include radio frequency identification (RFID) tags. This is a relatively simple extension in which each tag is associated with a particular object, yet it could drastically improve capabilities for a more complicated real-world setting. Finally, we plan to incorporate recent work in the area of *active learning* in an approach in which the most valuable episodes are chosen to be labeled first. This approach could model user abilities as well as system needs, in order to minimize useless, unlabeled, or mislabeled episodes.

REFERENCES

1. L. V. Ahn and L. Dabbish. Labeling images with a computer game. In *Proc. of CHI 2004*, pages 319 – 326, 2004.
2. L. F. Barrett and D. J. Barrett. An introduction to computerized experience sampling in psychology. *Social Science Computer Review*, 19(2):175–185, 2001.
3. J. Beaudin, S. Intille, and E. M. Tapia. Lessons learned using ubiquitous sensors for data collection in real homes. In *Extended Abstracts of CHI 2004*, pages 1359–1362, 2004.
4. C. Beckmann, S. Consolvo, and A. LaMarca. Some assembly required: Supporting end-user sensor installation in domestic ubiquitous computing environments. In *Proc. of UBICOMP 2004*, 2004.
5. S. Intille, E. M. Tapia, J. Rondoni, J. Beaudin, C. Kukla, S. Agarwal, and L. Bao. Tools for studying behavior and technology in natural settings. In *Proc. of UBICOMP 2003*, 2003.
6. J. Rondoni. Context-aware experience sampling for the design and study of ubiquitous technologies. Master's thesis, MIT, Sept. 2003.
7. D. Wilson and C. Atkeson. The narrator : A daily activity summarizer using simple sensors in an instrumented environment. In *Adjunct Proc. of UBICOMP 2003: Demonstrations*, pages 141 – 144, 2003.
8. D. Wilson and C. Atkeson. Automatic health monitoring using anonymous, binary sensors. In *CHI 2004 Workshops: Home Technologies to Keep Elders Connected*, 2004.